Internationale Standards zur Textauszeichnung (SGML, TEI)

Bader, Winfried

Internationale Standards zur Textauszeichnung (SGML, TEI)

Winfried Bader (Tübingen)*

Der Vortrag entstand in der Folge des Besuches eines Workshops der TEI *(Text Encoding Initiative)* im Dezember 1994 in Chicago, bei dem es darum ging, die Codierungsvorschläge der TEI unter dem Gesichtspunkt der Anwendung und Weitervermittlung kennenzulernen.

Bei der Texterfassung im Computer ist zu unterscheiden zwischen dem *character encoding* (Repräsentation der Einzelzeichen, bits und bytes), dem *page encoding* (Repräsentation der typographischen Realisierung einer Seite, z. B. PostScript) und dem *text encoding* (Repräsentation eines Textes mit seinen Strukturen und zusätzlichen interessanten Informationen). Textauszeichnung *(text encoding* oder *text markup)* ist eine Methode, eine oder mehrere Interpretationen des Textes explizit zu machen. Textauszeichnung ist immer eine Hinzufügung von inhaltlichem Wissen zu der bloßen Repräsentation der einzelnen Zeichen.

Die *Standard Generalized Markup Language* (SGML)

SGML ist ein Regelwerk zur Definition einer Textauszeichnungssprache mit der Grundidee, die Daten und die Verarbeitung strikt voneinander zu trennen. Das Ziel ist

- die Wiederverwendung von Texten in verschiedener Form und in verschiedenen Verarbeitungsgängen
- die Systemunabhängigkeit und Austauschbarkeit der Daten
- die Möglichkeit der Anreicherung von Texten mit Intelligenz (Zusatzinformationen, Interpretationen)
- eine wohlorganisierte Struktur und klar definierte Schnittstellen.

*** Protokoll des 63. Kolloquiums über die Anwendung der EDV in den Geisteswissenschaften an der Universität Tübingen am 11. Februar 1995.**

Die Dokument-Analyse

Da Textauszeichnung ein interpretativer Vorgang ist, muß ihr stets eine Dokument-Analyse vorausgehen. Dazu gehört zum einen das Untersuchen des Inhalts und der Struktur des Dokumentes, um die

- Komponenten des Dokumentes festzustellen und zu benennen
- ihr Verhältnis untereinander zu bestimmen
- ihre Eigenheiten zu erkennen,

und zum anderen das Abstecken der zusätzlich benötigten außertextlichen Informationen und Ziele hinsichtlich

- Herkunft, Quellenlage und situativem Kontext des Dokuments
- der organisatorischen Voraussetzungen für die Erfassung und Auszeichnung
- der technischen und inhaltlichen Ziele und Verwendungen des elektronischen Texts.

Bei der konkreten Auszeichnungsarbeit gibt es ein Abwägen zwischen zu viel (kostet Zeit, sichert aber die vielfältige Verwendung) und zu wenig (wichtige benötigte Information kann später fehlen) Auszeichnung. Als Richtlinie kann gelten: Die durch die Auszeichnung eingetragene Analyse des Textes muß wahr, hilfreich und handhabbar sein.

Die *Document Type Definition* (DTD)

Die in der Dokument-Analyse festgestellte Struktur eines Textes wird in der DTD nach den Regeln von SGML in Form eines Inhaltsmodells *(content model)* wiedergegeben. Das Dokument wird dabei vollständig in Form von hierarchischen Elementen beschrieben, beginnend beim größten Element (zugleich der Dokumenttyp) bis hin zu den einzelnen Zeichen, aus denen sich der Text zusammensetzt. Dabei wird für jedes Element vollständig mit Reihenfolge und Häufigkeit angegeben, aus welchen untergeordneten Elementen es besteht So besteht z. B. ein Element TEXT aus einem oder mehreren Elementen ÜBERSCHRIFT, denen jeweils ein oder mehrere Elemente ABSCHNITT folgen müssen. Eine ÜBERSCHRIFT besteht aus Daten (einzelnen Zeichen, Buchstaben), ein ABSCHNITT aus Daten oder Elementen HERVORHEBUNG, von denen jedes wiederum aus Daten besteht Damit ist ein einfacher Text komplett beschrieben. Man sieht daran zwei Dinge: Vor jeder Auszeichnung eines konkreten Textes muß die Erstellung einer entsprechenden DTD erfolgen. Andrerseits ist zu erwarten, das eine DTD auf viele gleichartige Texte zutrifft, so daß man für eine konkrete Textauszeichunng eine vorhandene, passende DTD anwenden kann. An diesem Punkt setzt die TEI an.

Die *Text Encoding Initiative* (TEI)

Die *Text Encoding Initiative* ist ein Projekt in dem sich 1987 die *Association for Computers and the Humanities* (ACH), die *Association for Computational*

Linguistics (ACL) und die *Association for Literary and Linguistics Computing* (ALLC) zusammengeschlossen haben, um ein Standardschema zur Textauszeichnung zu erarbeiten. Die Beratungen und Erarbeitungen, zu denen auch viele weitere Vereinigungen aus dem Gebiet der Text- und Bibliothekswissenschaft beitrugen, endeten mit der Veröffentlichung der ca. 1300 Seiten umfassenden Richtlinien *Guidelines for Electronic Text Encoding and Interchange (TEI P3)* im Mai 1994. Die Richtlinien sind in gedruckter Form oder auf CD-ROM als *DynaText edition* unter folgender Adresse zu beziehen: *TEI Orders, Oxford University Computing Services, 13 Banbury Road, Oxford 0X2 6NN, GB.* Außerdem sind die Richtlinien im *World Wide Web* unter dem URL *http -.lletext.Virginia.edulTEI.html* zugänglich.

Die Richtlinien der TEI sind im Kern SGML-konforme DTDs. Vorgefertigte DTDs werden bereitgestellt, so daß Anwenderinnen - nach einer eingehenden eigenen Dokument-Analyse - die passenden Elemente direkt auf ihren Text anwenden können, ohne mühsam sich selbst um die notwendigen Definitionen zu sorgen. Texte, die nach den Richtlinien der TEI ausgezeichnet sind, sind SGML-konforme Dokumente, die sich mit jeder SGML-Software bearbeiten lassen. Der entscheidende Vorteil in der Anlage der TEI-DTDs ist ihr modularer Aufbau. Eine allgemeine DTD, die im Kern auf alle Texte zutrifft, kann je nach Textart (Prosa, Poesie, Drama, gesprochene Sprache) und Ziel der Erfassung (Textkritik, linguistische Analyse) um spezielle DTDs erweitert werden.

Obligat für alle Texte, die entsprechend den DTDs der TEI ausgezeichnet werden, ist der TEI-Header. Er enthält die »bibliographische« Information über das betreffende elektronische Dokument. Es ist die elektronische »Titelseite«. Der Header enthält Informationen über den Inhalt des Dokumentes, die (gedruckte/handgeschriebene) Vorlage, die Codierungskonventionen, die Verfügbarkeit, die Verantwortlichen und die Korrekturen. Der Header dient dazu, das Dokument eindeutig identifizierbar zu machen. Für die Archivierung und Dokumentation wird der/die Bibliothekar/in hieraus die notwendigen Informationen zur Erstellung des »Karteikärtchens« über dieses elektronische Dokument nehmen. Das Dokument wird durch den Header zitierfähig (und damit auch der/diejenige, der/die den Text in die elektronische Form gebracht und ausgezeichnet hat).

Beispiel

Als konkretes Beispiel der Auszeichnung eines Textes soll das Anbringen eines Querverweises dienen. Das Beispiel zeigt das *tagging* der Elemente mit Anfangs- und Endekennung und die Möglichkeit, einem Element im Anfangs-tag Attribute zuzuschreiben. Für den Querverweis muß das Element, auf das verwiesen wird, eine eindeutige ID haben, die als Attribut bei jedem Element möglich ist *(<div3 id=kap070103>).* Damit kann durch Angabe dieser ID mit einem Attribut des Typs IDREF von einem anderen Element *(<ref target=kap070103>)* darauf verwiesen werden.

<head>2.5 Der Querverweismechanismus<lhead>
<p> ... Siehe unten in <ref target=kap070103>Kapitel 7.1.3</reß>
<— Weiter unten im Dokument —>
<div3 id=kap070103>
<head>7.1.3 ....

Die Verbindung zwischen den beiden Textstellen ist damit eindeutig hergestellt. Es liegt nun an der Verarbeitungssoftware, ob dieser Querverweis beim Anschauen *(browsen)* wie ein Hyperlink benutzt werden kann. Für die Auszeichnung von »Hyperlinks« stellt die TEI im Grunde die gleichen, aber viel reichhaltigeren Möglichkeiten zur Verfügung wie das bekannte HTML *(Hyper Text Markup Language).* Es hegt an der Einfachheit und Beschränktheit der sonstigen Möglichkeiten von HTML, daß es dafür bereits weitverbreitete Software gibt, die ein Browsen unter Benutzung der Hyperlinks zufriedenstellend ermöglicht Für die komplizierteren SGML-DTDs, wie sie die TEI bereitstellt, ist die Software noch nicht soweit. Für die wissenschaftliche Auszeichnung von Texten ist die HTML-DTD jedoch ungeeignet. HTML und TEI verhalten sich zueinander etwa wie Tageszeitung und historisch-kritische Edition.

Softwareanforderungen für SGML-Dokumente

SGML ist eine deskriptive Textauszeichnung. Sie hat direkt nichts mit der Weiterverarbeitung zu tun. SGML-Dokumente können auf verschiedene Weise weiterverarbeitet und weiterverwendet werden, wofür es zahlreiche Software gibt. Nicht zuletzt kann TUSTEP vorzüglich für die Weiterverarbeitung von SGML-Dokumenten eingesetzt werden. Echte SGML Software ist in der Lage, SGML-konforme DTDs einzulesen, diese zu verstehen und damit dieser DTD entsprechende Dokumente zu verarbeiten. Die Software zur Verarbeitung von SGML-Dokumenten kann in verschiedene Aufgabenbereiche eingeteilt werden:

Checker, Parser. Programme, die die Syntax und die Konformität eines SGML-Dokumentes bezüglich einer DTD abprüfen.

Editor. Unterstützung beim Schreiben von *tags;* Umsetzung einer DTD in vorgegebene Strukturen zur Texterfassung.

Konvertierer. Konvertierung von bestimmten Datenformaten (Textverarbeitung) in SGML-Dokumente.

Retrieval: Suchen und Recherchieren unter Einbeziehung der ausgezeichneten Informationen.

Browser: Anzeigen von SGML-Dokumenten am Bildschirm unter typographischer Auswertung der *tags.*

Formatter. Satz und Druck von SGML-Dokumenten in einem festzulegenden Stil; Umsetzung von *tags* in typographische Anweisungen und Unterdrükkung von bestimmten Elementen.

FORTHCOMING EVENTS

ZHSF-HERB STSEMIN ARE 1996
»METHODIK DER HISTORISCHEN SOZIALFORSCHUNG«
- GRUNDKURS -
Köln, 31. August bis 14. September 1996

Kursangebot: Vierzehntägiger Einführungskurs in die wissenschaftstheoretischen, methodologischen, forschungstechnischen und statistischen Grundlagen der Historischen Sozialforschung und ihrer EDV-Anwendung.

Inhalte: Grundlagen der Methodik Historischer Sozialforschung: Theorie-/Hypothesenbildung; die 'empirische Übersetzung' von Forschungsproblemen (Problemformulierung, Operationalisierung, Indikatoren); Auswahl der Untersuchungseinheiten (Grundlagen der Stichprobenbildung, Auswahlverfahren); Erhebungsverfahren; computergestützte Datenerfassung und -aufbereitung; Datenanalyse als Anwendung statistischer Modelle und Verfahren: Grundlagen der Statistik; univariate und bivariate Häufigkeitsverteilungen, statistische Maßzahlen von Häufigkeitsverteilungen, Quantifizierung von Zusammenhängen in zwei- und dreidimensionale Häufigkeitstabellen, bivariate Regressions- und Varianzanalyse; Ausblick auf weiterführende Datenanalyseverfahren; EDV-Einsatz: Einführung in die Arbeit mit dem Personalcomputer, Anwendung des Statistik-Programmpakets SPSS/PC für Windows.

Vermittlung: Durch die Dozenten erfolgt eine Einführung in die Lerninhalte. Die Kursteilnehmer sollen die Lerninhalte am Beispiel von historischen Quellen und eines Datensatzes aus der Historischen Sozialforschung in Arbeitsgruppen forschungspraktisch umsetzen, die Übungsdatensätze selbständig unter Einsatz von EDV auswerten und die Analyseergebnisse mit Hilfe eines Textverarbeitungssystems schriftlich darstellen. Während der gesamten Kurslaufzeit stehen die Dozenten und ZHSF-Mitarbeiter für die Beratung der Teilnehmer bei der Durchführung eigener Forschungsarbeiten zur Verfügung. Darüberhinaus besteht die Möglichkeit, in Teilnehmervorträgen solche Forschungsarbeiten vor dem Plenum zu diskutieren.

Vorkenntnisse: Im Hinblick auf die Methodik der Historischen Sozialforschung und deren EDV-Anwendung werden bei den Teilnehmern des Grundkurses keine spezifischen Vorkenntnisse erwartet.

Teilnehmerkreis: Der Grundkurs ist für graduierte Teilnehmer in Forschung

und Lehre sowie für fortgeschrittene Studenten am Ende ihres Studiums eingerichtet.

Dozenten: Dr. Edwin Keiner (Universität Frankfurt), Dipl.-Wirtsch. Karl Pierau (ZA-ZHSF), PD Dr. Johann Bacher (Universität Linz), PD Dr. Wilhelm H. Schröder (ZA-ZHSF).

Ort/Termin: 31.8.-14.9.96 in den Seminarräumen und den PC-Pools der Universität Köln.

Kursgebühren: DM 120,- (Studenten: DM 60,-).

Kursleitung: PD Dr. Wilhelm H. Schröder.

Sekretariat: Lilo Montes.

EDV-Org.: Rainer Hinterberg.

Anmeldung: Bitte fordern Sie die Anmeldeunterlagen an beim Zentralarchiv für Empirische Sozialforschung, Abteilung Zentrum für Historische Sozialforschung, Lilienconstr. 6, 50931 Köln (Tel.: 0221/47694-34; Fax: 0221/47694-55).

Anmeldeschluß: 15. Juni 1996.

ZHSF-HERBSTSEMINARE 1996
»METHODIK DER HISTORISCHEN SOZIALFORSCHUNG«
- AUFBAUKURS -
Köln, 31. August bis 14. September 1996

Kursangebot: Vierzehntägiger Weiterbildungskurs für Fortgeschrittene zu den Grundlagen des allgemeinen linearen Regressionsmodells (ALM) sowie ausgewählter Submodelle des ALM und ihrer EDV-Anwendung.

Inhalte: *Repetitorium:* Grundlagen der Statistik; uni- und bivariate Datenanalyse; Konzept der statistischen Kontrolle von Drittvariablen; computergestützte Datenanalyse mit dem Personalcomputer unter Anwendung des Statistik-Programmpakets SPSS/PC für Windows (Crashkurs für das Themenspektrum des Grundkurses).

Kernkurs: Multiple Regressionsanalyse; Konzepte der Inferenzstatistik; mehrfache Varianzanalyse; Regressionsanalyse mit Dummy-Variablen; Analysemodelle mit diskreten abhängigen Merkmalen (lineares Wahrscheinlichkeits- und logistisches Regressionsmodell); Ausblick auf spezielle Analysemodelle: Erklärung von 'Ereignissen im Zeitverlauf (Fragestellung der Ereignisdatenanalyse) und Berücksichtigung von 'Zeit und Dynamik' (grundlegende Fragestellungen der Zeitreihenanalyse); EDV-Einsatz: Anwendung von SPSS/PC für Windows.

Vermittlung: Durch die Dozenten erfolgt eine Einführung in die Lerninhalte. Die Kursteilnehmer sollen die Lerninhalte am Beispiel von histori-

sehen Quellen und eines Datensatzes aus der Historischen Sozialforschung in Arbeitsgruppen forschungspraktisch umsetzen, die Übungsdatensätze selbständig unter Einsatz von EDV auswerten und die Analyseergebnisse mit Hilfe eines Textverarbeitungssystem schriftlich darstellen. Während der gesamten Kurslaufzeit stehen die Dozenten und ZHSF-Mitarbeiter für die Beratung der Teilnehmer bei der Diffchführung eigener Forschungsarbeiten zur Verfügung. Darüberhinaus besteht die Möglichkeit, in Teilnehmervorträgen solche Forschungsarbeiten vor dem Plenum zu diskutieren.

Vorkenntnisse: Es ist die vorhergehende Teilnahme an einem ZHSF-Grundkurs erforderlich bzw. werden zumindest vergleichbare Grundkenntnisse in der Methodik der Historischen Sozialforschung, der Planung und Durchführung empirischer Forschung, der uni- und bivariaten Datenanalyse (d.h. statistische Maßzahlen, Zusammenhangs- und Abhängigkeitsanalyse in bivariaten Häufigkeitstabellen, bivariate Regressions- und Varianzanalyse) und der Personalcomputer-Anwendung (SPSS/PC für Windows) vorausgesetzt.

Teilnehmerkreis: Der Aufbaukurs ist für graduierte Teilnehmer in Forschung und Lehre sowie für fortgeschrittene Studenten am Ende ihres Studiums eingerichtet.

Dozenten: Dipl. Soz. Jürgen Sensen (ZA-ZHSF), Dipl. Volksw. Dieter Ohr (Universität zu Köln), Hermann Dülmer M.A. (Universität zu Köln), PD Dr. Rainer Metz (ZA-ZHSF).

Ort/Termin: 31.8.-14.9.96 in den Seminarräumen und den PC-Pools der Universiät Köln.

Kursgebühren: DM 120,- (Studenten: DM 60,-).

Kursleitung: PD Dr. Wilhelm H. Schröder.

Sekretariat: Lilo Montes.

EDV-Org.: Ralph Ponemereo.

Anmeldung: Bitte fordern Sie die Anmeldeunterlagen an beim Zentralarchiv für Empirische Sozialforschung, Abteilung Zentrum für Historische Sozialforschung, Liliencronstr. 6, 50931 Köln (Tel.: 0221 / 47694-34; Fax: 0221 /47694-55).

Anmeldeschluß: 15. Juni 1996.

XI International Conference of the Association for History and Computing
Data Modelling, Modelling History

Moscow, August 21-24, 1996

After several decades of experience with computer-assisted processing of various kinds of historical data, historical computing has consolidated its status as a mature empirical (auxiliary) discipline. Further advance of historical computing should offer a firm conceptual and theoretical framework. This requires a further development of IT applications with advanced data modelling techniques and methods of computer-assisted modelling of historical processes. Such models seem to be powerful tools for connecting data and theory, but they also make demands on the nature of the theory and data. Should models be theory-driven or data-driven? Should a projects analytical aims, its software, and its data modelling be considered independently? What are particular features of historical data which make it different from other data processed by computers?

The XI AHC Conference will offer a platform to discuss modelling and processing of different kinds of historical data - from census-like data to free text and images - as well as advanced analytical tools in the broader context of continuous computer-aided historical research, starting with the correct treatment of digitalized sources, arriving at theoretically grounded and convincing research results.

Sub-Themes

I. Historical Data Modelling
- historical text models
- historical software benchmark
- new database models/techniques
- modelling highly structured historical data
- models of historical images
- fuzzy models of historical data

II. Modelling History
- statistical models of historical structures
- modelling dynamics of historical processes
- event history analysis
- modelling of spatial historical structures
- simulation of processes in the past
- the *pros* and *cons* of counterfactual modelling
- statistical text analysis: benefits and limitations for historical research
- artificial intelligence models in historical research

- simulation in computer-assisted teaching of history
- multimedia presentation of historical evidence
- history on the Internet: a model of information exchange in the professional community

III. Towards theory of historical computing: the bridge between historical data modelling and analysis

Organizing committee: Dr. Leonid Borodkin (Moscow); Dr. Segey Kashchenko (St.- Petersburg); Dr. Vladimir Vladimirov (Barnaul); Dr. Irina Garskova (Moscow); Dr. Dilyara Ibragimova (Moscow); Mg Valéry Lazarev (Moscow).

International Programme Committee: Dr. Onno Boonstra (Nijmegen); Dr. Leonid Borodkin (Moscow); Dr. Peter Denley (London); Dr. Peter Doom (Leiden); Dr. Stefan Fogelvik (Stockholm); Dr. José Igartua (Montreal); Dr. Jan Oldervoll (Bergen); Dr. Manfred Thaller (Goettingen).

This conference is organized by the Association for History & Computing in cooperation with the Faculty of History, Moscow State University, Russian Academy of Sciences.

Call for papers:
Proposals are invited for sessions, papers and other contributions for main sessions (30 min. reading time), special sessions (20 min.), project presentations (10 min.) and demonstrations, all related to one of the mentioned sub-themes.

Conference language will be English.

For further information please contact:

AHC96
Lab for Historical Information Science
Faculty of History
Moscow State University
Vorobyevy Hills
119899 Moscow/Russia
TeVFax: (095) 939-11-65
E-mail: ahc96@histchem.msu.su